Chapter Goals

- A taxonomy of fields relevant to learning from data, including data science, machine learning, and statistics.
- Data mining life-cycle

Lab Goals

- Hands-on familiarity with weka, pandas, and matplotlib.

1.1 & 1.5 Data Mining, Machine Learning, and Statistics

Task 1.1.1 Select three courses from your studied curriculum whose methodology is based on Math/Logic, Science, and Engineering, respectively. Illustrate your decisions.

Task 1.1.2 Can we learn from data without using any machine learning algorithm? Draw some examples.

Task 1.1.3 Judge whether the above data has a pattern or a special feature, and explain.

```
# Data 1
[
    [1, 10],
    [2, 9],
    [3, 8],
    [4, 7],
    [5, 6],
]

# Data 2
[1.1, 1.2, 0.8, 0.9, 1, 1,1]

# Data 3
[1, 4, 9, 16, 25, 36]

# Data 4
[19, 35] # The time in which the author wrote line
```

Answer

Task 1.1.4 Construct a linear model approximating *Data 1* without using any machine learning algorithm. You can use the below code snippet

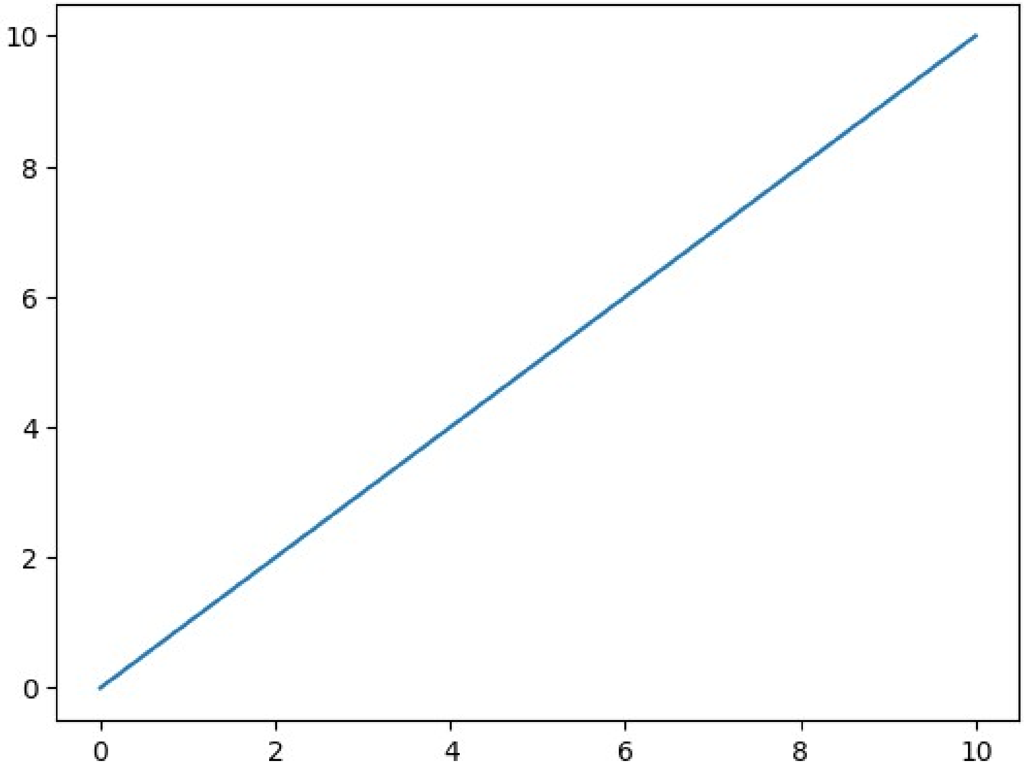
```
# Linear model visualized

# Define the linear equation: y = mx + b
m = 1.00  # Slope
b = 0.00  # Y-intercept

# Generate x-values and y-values using the equation
x_values = np.linspace(0, 10)  # Adjust the range and number of points
as needed
y_values = (m * x_values) + b

# Plot the scatter and line
plt.plot(x_values, y_values)

[<matplotlib.lines.Line2D at 0x7908962ca740>]
```



Challenge 1.1.5 Research traditional statistics methodologies, through which statisticians were able to construct models and verify their accuracy.

1.2. Simple Examples: The Weather Problem and Others (Hello-world Tutorials)

Modules & Datasets Setup

```
# @title
!apt-get install default-jdk
!apt install libgraphviz-dev
```

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  default-jdk-headless default-jre default-jre-headless fonts-dejavu-
core
  fonts-dejavu-extra libatk-wrapper-java libatk-wrapper-java-jni
libfontenc1
  libice-dev libsm-dev libxkbfile1 libxt-dev libxtst6 libxxf86dga1
  openjdk-11-jdk openjdk-11-jre x11-utils
Suggested packages:
  libice-doc libsm-doc libxt-doc openjdk-11-demo openjdk-11-source
visualvm
  mesa-utils
The following NEW packages will be installed:
  default-jdk default-jdk-headless default-jre default-jre-headless
  fonts-dejavu-core fonts-dejavu-extra libatk-wrapper-java
  libatk-wrapper-java-jni libfontenc1 libice-dev libsm-dev libxkbfile1
  libxt-dev libxtst6 libxxf86dga1 openjdk-11-jdk openjdk-11-jre x11-
utils
0 upgraded, 18 newly installed, 0 to remove and 18 not upgraded.
Need to get 5,518 kB of archives.
After this operation, 15.8 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/main amd64 default-jre-
headless amd64 2:1.11-72build2 [3,042 B]
Get:2 http://archive.ubuntu.com/ubuntu jammy/main amd64 libxtst6 amd64
2:1.2.3-1build4 [13.4 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64
openjdk-11-jre amd64 11.0.20.1+1-0ubuntu1~22.04 [213 kB]
Get:4 http://archive.ubuntu.com/ubuntu jammy/main amd64 default-jre
amd64 2:1.11-72build2 [896 B]
Get:5 http://archive.ubuntu.com/ubuntu jammy/main amd64 default-jdk-
headless amd64 2:1.11-72build2 [942 B]
Get:6 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64
openjdk-11-jdk amd64 11.0.20.1+1-0ubuntu1~22.04 [1,331 kB]
Get:7 http://archive.ubuntu.com/ubuntu jammy/main amd64 default-jdk
amd64 2:1.11-72build2 [908 B]
Get:8 http://archive.ubuntu.com/ubuntu jammy/main amd64 fonts-dejavu-
core all 2.37-2build1 [1,041 kB]
```

```
Get:9 http://archive.ubuntu.com/ubuntu jammy/main amd64 fonts-dejavu-
extra all 2.37-2build1 [2,041 kB]
Get:10 http://archive.ubuntu.com/ubuntu jammy/main amd64 libfontenc1
amd64 1:1.1.4-1build3 [14.7 kB]
Get:11 http://archive.ubuntu.com/ubuntu jammy/main amd64 libxkbfile1
amd64 1:1.1.0-1build3 [71.8 kB]
Get:12 http://archive.ubuntu.com/ubuntu jammy/main amd64 libxxf86dga1
amd64 2:1.1.5-0ubuntu3 [12.6 kB]
Get:13 http://archive.ubuntu.com/ubuntu jammy/main amd64 x11-utils
amd64 7.7+5build2 [206 kB]
Get:14 http://archive.ubuntu.com/ubuntu jammy/main amd64 libatk-
wrapper-java all 0.38.0-5build1 [53.1 kB]
Get:15 http://archive.ubuntu.com/ubuntu jammy/main amd64 libatk-
wrapper-java-jni amd64 0.38.0-5build1 [49.0 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy/main amd64 libice-dev
amd64 2:1.0.10-1build2 [51.4 kB]
Get:17 http://archive.ubuntu.com/ubuntu jammy/main amd64 libsm-dev
amd64 2:1.2.3-1build2 [18.1 kB]
Get:18 http://archive.ubuntu.com/ubuntu jammy/main amd64 libxt-dev
amd64 1:1.2.1-1 [396 kB]
Fetched 5,518 kB in 3s (2,049 kB/s)
Selecting previously unselected package default-jre-headless.
(Reading database ... 120895 files and directories currently
installed.)
Preparing to unpack .../00-default-jre-headless_2%3a1.11-
72build2_amd64.deb ...
Unpacking default-jre-headless (2:1.11-72build2) ...
Selecting previously unselected package libxtst6:amd64.
Preparing to unpack .../01-libxtst6_2%3a1.2.3-1build4_amd64.deb ...
Unpacking libxtst6:amd64 (2:1.2.3-1build4) ...
Selecting previously unselected package openjdk-11-jre:amd64.
Preparing to unpack .../02-openjdk-11-jre_11.0.20.1+1-
0ubuntu1~22.04_amd64.deb ...
Unpacking openjdk-11-jre:amd64 (11.0.20.1+1-0ubuntu1~22.04) ...
Selecting previously unselected package default-jre.
Preparing to unpack .../03-default-jre_2%3a1.11-72build2_amd64.deb ...
Unpacking default-jre (2:1.11-72build2) ...
Selecting previously unselected package default-jdk-headless.
Preparing to unpack .../04-default-jdk-headless_2%3a1.11-
72build2_amd64.deb ...
Unpacking default-jdk-headless (2:1.11-72build2) ...
Selecting previously unselected package openjdk-11-jdk:amd64.
Preparing to unpack .../05-openjdk-11-jdk_11.0.20.1+1-
0ubuntu1~22.04_amd64.deb ...
Unpacking openjdk-11-jdk:amd64 (11.0.20.1+1-0ubuntu1~22.04) ...
Selecting previously unselected package default-jdk.
Preparing to unpack .../06-default-jdk_2%3a1.11-72build2_amd64.deb ...
Unpacking default-jdk (2:1.11-72build2) ...
Selecting previously unselected package fonts-dejavu-core.
```

```
Preparing to unpack .../07-fonts-dejavu-core_2.37-2build1_all.deb ...
Unpacking fonts-dejavu-core (2.37-2build1) ...
Selecting previously unselected package fonts-dejavu-extra.
Preparing to unpack .../08-fonts-dejavu-extra_2.37-2build1_all.deb ...
Unpacking fonts-dejavu-extra (2.37-2build1) ...
Selecting previously unselected package libfontenc1:amd64.
Preparing to unpack .../09-libfontenc1_1%3a1.1.4-1build3_amd64.deb ...
Unpacking libfontenc1:amd64 (1:1.1.4-1build3) ...
Selecting previously unselected package libxkbfile1:amd64.
Preparing to unpack .../10-libxkbfile1_1%3a1.1.0-1build3_amd64.deb ...
Unpacking libxkbfile1:amd64 (1:1.1.0-1build3) ...
Selecting previously unselected package libxxf86dga1:amd64.
Preparing to unpack .../11-libxxf86dga1_2%3a1.1.5-
0ubuntu3_amd64.deb ...
Unpacking libxxf86dga1:amd64 (2:1.1.5-0ubuntu3) ...
Selecting previously unselected package x11-utils.
Preparing to unpack .../12-x11-utils_7.7+5build2_amd64.deb ...
Unpacking x11-utils (7.7+5build2) ...
Selecting previously unselected package libatk-wrapper-java.
Preparing to unpack .../13-libatk-wrapper-java_0.38.0-
5build1_all.deb ...
Unpacking libatk-wrapper-java (0.38.0-5build1) ...
Selecting previously unselected package libatk-wrapper-java-jni:amd64.
Preparing to unpack .../14-libatk-wrapper-java-jni_0.38.0-
5build1_amd64.deb ...
Unpacking libatk-wrapper-java-jni:amd64 (0.38.0-5build1) ...
Selecting previously unselected package libice-dev:amd64.
Preparing to unpack .../15-libice-dev_2%3a1.0.10-1build2_amd64.deb ...
Unpacking libice-dev:amd64 (2:1.0.10-1build2) ...
Selecting previously unselected package libsm-dev:amd64.
Preparing to unpack .../16-libsm-dev_2%3a1.2.3-1build2_amd64.deb ...
Unpacking libsm-dev:amd64 (2:1.2.3-1build2) ...
Selecting previously unselected package libxt-dev:amd64.
Preparing to unpack .../17-libxt-dev_1%3a1.2.1-1_amd64.deb ...
Unpacking libxt-dev:amd64 (1:1.2.1-1) ...
Setting up default-jre-headless (2:1.11-72build2) ...
Setting up libice-dev:amd64 (2:1.0.10-1build2) ...
Setting up libsm-dev:amd64 (2:1.2.3-1build2) ...
Setting up libxtst6:amd64 (2:1.2.3-1build4) ...
Setting up libxxf86dga1:amd64 (2:1.1.5-0ubuntu3) ...
Setting up openjdk-11-jre:amd64 (11.0.20.1+1-0ubuntu1~22.04) ...
Setting up default-jre (2:1.11-72build2) ...
Setting up libfontenc1:amd64 (1:1.1.4-1build3) ...
Setting up default-jdk-headless (2:1.11-72build2) ...
Setting up libxt-dev:amd64 (1:1.2.1-1) ...
Setting up fonts-dejavu-core (2.37-2build1) ...
Setting up fonts-dejavu-extra (2.37-2build1) ...
Setting up openjdk-11-jdk:amd64 (11.0.20.1+1-0ubuntu1~22.04) ...
update-alternatives: using
```

```
/usr/lib/jvm/java-11-openjdk-amd64/bin/jconsole to provide
/usr/bin/jconsole (jconsole) in auto mode
Setting up libxkbfile1:amd64 (1:1.1.0-1build3) ...
Setting up default-jdk (2:1.11-72build2) ...
Setting up x11-utils (7.7+5build2) ...
Setting up libatk-wrapper-java (0.38.0-5build1) ...
Setting up libatk-wrapper-java-jni:amd64 (0.38.0-5build1) ...
Processing triggers for hicolor-icon-theme (0.17-2) ...
Processing triggers for libc-bin (2.35-0ubuntu3.1) ...
/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_0.so.3 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic
link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic
link

Processing triggers for man-db (2.10.2-1) ...
Processing triggers for fontconfig (2.13.1-4.2ubuntu5) ...
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  libgail-common libgail18 libgtk2.0-0 libgtk2.0-bin libgtk2.0-common
  libgvc6-plugins-gtk librsvg2-common libxdot4
Suggested packages:
  gvfs
The following NEW packages will be installed:
  libgail-common libgail18 libgraphviz-dev libgtk2.0-0 libgtk2.0-bin
  libgtk2.0-common libgvc6-plugins-gtk librsvg2-common libxdot4
0 upgraded, 9 newly installed, 0 to remove and 18 not upgraded.
Need to get 2,433 kB of archives.
After this operation, 7,694 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/main amd64 libgtk2.0-
common all 2.24.33-2ubuntu2 [125 kB]
Get:2 http://archive.ubuntu.com/ubuntu jammy/main amd64 libgtk2.0-0
amd64 2.24.33-2ubuntu2 [2,037 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy/main amd64 libgail18
amd64 2.24.33-2ubuntu2 [15.9 kB]
Get:4 http://archive.ubuntu.com/ubuntu jammy/main amd64 libgail-common
```

```
amd64 2.24.33-2ubuntu2 [132 kB]
Get:5 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libxdot4
amd64 2.42.2-6 [16.4 kB]
Get:6 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libgvc6-
plugins-gtk amd64 2.42.2-6 [22.6 kB]
Get:7 http://archive.ubuntu.com/ubuntu jammy/universe amd64
libgraphviz-dev amd64 2.42.2-6 [58.5 kB]
Get:8 http://archive.ubuntu.com/ubuntu jammy/main amd64 libgtk2.0-bin
amd64 2.24.33-2ubuntu2 [7,932 B]
Get:9 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64
librsvg2-common amd64 2.52.5+dfsg-3ubuntu0.2 [17.7 kB]
Fetched 2,433 kB in 2s (1,061 kB/s)
Selecting previously unselected package libgtk2.0-common.
(Reading database ... 121406 files and directories currently
installed.)
Preparing to unpack .../0-libgtk2.0-common_2.24.33-
2ubuntu2_all.deb ...
Unpacking libgtk2.0-common (2.24.33-2ubuntu2) ...
Selecting previously unselected package libgtk2.0-0:amd64.
Preparing to unpack .../1-libgtk2.0-0_2.24.33-2ubuntu2_amd64.deb ...
Unpacking libgtk2.0-0:amd64 (2.24.33-2ubuntu2) ...
Selecting previously unselected package libgail18:amd64.
Preparing to unpack .../2-libgail18_2.24.33-2ubuntu2_amd64.deb ...
Unpacking libgail18:amd64 (2.24.33-2ubuntu2) ...
Selecting previously unselected package libgail-common:amd64.
Preparing to unpack .../3-libgail-common_2.24.33-
2ubuntu2_amd64.deb ...
Unpacking libgail-common:amd64 (2.24.33-2ubuntu2) ...
Selecting previously unselected package libxdot4:amd64.
Preparing to unpack .../4-libxdot4_2.42.2-6_amd64.deb ...
Unpacking libxdot4:amd64 (2.42.2-6) ...
Selecting previously unselected package libgvc6-plugins-gtk.
Preparing to unpack .../5-libgvc6-plugins-gtk_2.42.2-6_amd64.deb ...
Unpacking libgvc6-plugins-gtk (2.42.2-6) ...
Selecting previously unselected package libgraphviz-dev:amd64.
Preparing to unpack .../6-libgraphviz-dev_2.42.2-6_amd64.deb ...
Unpacking libgraphviz-dev:amd64 (2.42.2-6) ...
Selecting previously unselected package libgtk2.0-bin.
Preparing to unpack .../7-libgtk2.0-bin_2.24.33-2ubuntu2_amd64.deb ...
Unpacking libgtk2.0-bin (2.24.33-2ubuntu2) ...
Selecting previously unselected package librsvg2-common:amd64.
Preparing to unpack .../8-librsvg2-common_2.52.5+dfsg-
3ubuntu0.2_amd64.deb ...
Unpacking librsvg2-common:amd64 (2.52.5+dfsg-3ubuntu0.2) ...
Setting up libxdot4:amd64 (2.42.2-6) ...
Setting up librsvg2-common:amd64 (2.52.5+dfsg-3ubuntu0.2) ...
Setting up libgtk2.0-common (2.24.33-2ubuntu2) ...
Setting up libgtk2.0-0:amd64 (2.24.33-2ubuntu2) ...
Setting up libgvc6-plugins-gtk (2.42.2-6) ...
Setting up libgail18:amd64 (2.24.33-2ubuntu2) ...
Setting up libgtk2.0-bin (2.24.33-2ubuntu2) ...
Setting up libgail-common:amd64 (2.24.33-2ubuntu2) ...
Setting up libgraphviz-dev:amd64 (2.42.2-6) ...
Processing triggers for libc-bin (2.35-0ubuntu3.1) ...
/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_0.so.3 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic
link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a
symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic
link

Processing triggers for man-db (2.10.2-1) ...
Processing triggers for libgdk-pixbuf-2.0-0:amd64 (2.42.8+dfsg-
1ubuntu0.2) ...
```

```
# @title
!pip install pygraphviz
!pip install python-javabridge
!pip install python-weka-wrapper3
!pip install sklearn-weka-plugin
```

```
Collecting pygraphviz
  Downloading pygraphviz-1.11.zip (120 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0.0/120.8 kB ? eta -:--:--
━━━━━━━━━━━━━━━━━━━━ ━━━━━━━━━━━━━━━━━━━ 61.4/120.8 kB 1.7 MB/s eta
0:00:01 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 120.8/120.8 kB 2.4
MB/s eta 0:00:00
etadata (setup.py) ... e=pygraphviz-1.11-cp310-cp310-linux_x86_64.whl
size=175927
sha256=65521de22428a5c41afd460979d8dfe5cc868efdd899874bc4ba5d6f26e00df
6
  Stored in directory:
/root/.cache/pip/wheels/5b/ee/36/f47a0d35664fbe1a2b5a433ae33c6ad636b00
bb231f68a9aaa
Successfully built pygraphviz
Installing collected packages: pygraphviz
Successfully installed pygraphviz-1.11
Collecting python-javabridge
```

```
  Downloading python-javabridge-4.0.3.tar.gz (1.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.3/1.3 MB 14.4 MB/s eta
0:00:00
etadata (setup.py) ... ent already satisfied: numpy>=1.20.1 in
/usr/local/lib/python3.10/dist-packages (from python-javabridge)
(1.23.5)
Building wheels for collected packages: python-javabridge
  Building wheel for python-javabridge (setup.py) ...
e=python_javabridge-4.0.3-cp310-cp310-linux_x86_64.whl size=1743153
sha256=3e3bb4a4f4df84a11683b01645eb735b7d70eced92c667b09215385df8691d2
c
  Stored in directory:
/root/.cache/pip/wheels/35/58/be/c5d71b71a9dd6585f897fa5b2d021e03962eb
30d6b20797396
Successfully built python-javabridge
Installing collected packages: python-javabridge
Successfully installed python-javabridge-4.0.3
Collecting python-weka-wrapper3
  Downloading python-weka-wrapper3-0.2.14.tar.gz (15.9 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 15.9/15.9 MB 3.3 MB/s eta
0:00:00
etadata (setup.py) ... ent already satisfied: python-javabridge>=4.0.0
in /usr/local/lib/python3.10/dist-packages (from python-weka-wrapper3)
(4.0.3)
Requirement already satisfied: numpy in
/usr/local/lib/python3.10/dist-packages (from python-weka-wrapper3)
(1.23.5)
Requirement already satisfied: packaging in
/usr/local/lib/python3.10/dist-packages (from python-weka-wrapper3)
(23.1)
Collecting configurable-objects (from python-weka-wrapper3)
  Downloading configurable-objects-0.0.1.tar.gz (4.4 kB)
  Preparing metadata (setup.py) ... ple-data-flow (from python-weka-
wrapper3)
  Downloading simple-data-flow-0.0.1.tar.gz (16 kB)
  Preparing metadata (setup.py) ... ple-data-flow
  Building wheel for python-weka-wrapper3 (setup.py) ...
e=python_weka_wrapper3-0.2.14-py3-none-any.whl size=14496261
sha256=04debd890fbb24010c3461396d2ce492b7f5d06de576555276e1edc6eff1dc8
8
  Stored in directory:
/root/.cache/pip/wheels/80/c5/f2/412fa8d3b181151e11b68d46daa52f96e9b83
2a2eca4bc6c88
  Building wheel for configurable-objects (setup.py) ...
e=configurable_objects-0.0.1-py3-none-any.whl size=4695
sha256=27d34557fbc11dc450b540a6576acb1fdeb64ddbae3a2c6c4b7c759905a1dc3
6
  Stored in directory:
/root/.cache/pip/wheels/ef/11/bc/75ac8b0592c38dc42412942c37d3947faf0b2
```

```
22bad150132a1
  Building wheel for simple-data-flow (setup.py) ... ple-data-flow: 
filename=simple_data_flow-0.0.1-py3-none-any.whl size=19063 
sha256=f267326b7eb2749907b7f6c9ff12716ba813f766808a182f46b79a2e10f8a68
f
  Stored in directory: 
/root/.cache/pip/wheels/b3/02/23/4aec0db3dae7152dd268d6de385905116af55
229c1a8e81303
Successfully built python-weka-wrapper3 configurable-objects simple-
data-flow
Installing collected packages: configurable-objects, simple-data-flow,
python-weka-wrapper3
Successfully installed configurable-objects-0.0.1 python-weka-
wrapper3-0.2.14 simple-data-flow-0.0.1
Collecting sklearn-weka-plugin
  Downloading sklearn-weka-plugin-0.0.7.tar.gz (69 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 69.8/69.8 kB 1.6 MB/s eta 
0:00:00
etadata (setup.py) ... ent already satisfied: numpy in 
/usr/local/lib/python3.10/dist-packages (from sklearn-weka-plugin) 
(1.23.5)
Requirement already satisfied: python-weka-wrapper3>=0.2.5 in 
/usr/local/lib/python3.10/dist-packages (from sklearn-weka-plugin) 
(0.2.14)
Collecting sklearn (from sklearn-weka-plugin)
  Downloading sklearn-0.0.post9.tar.gz (3.6 kB)
  error: subprocess-exited-with-error
  
  × python setup.py egg_info did not run successfully.
  │ exit code: 1
  ╰─> See above for output.
  
  note: This error originates from a subprocess, and is likely not a 
problem with pip.
  Preparing metadata (setup.py) ... error: metadata-generation-failed

× Encountered error while generating package metadata.
╰─> See above for output.

note: This is an issue with the package mentioned above, not pip.
hint: See above for details.

# @title
#Restart runtime after installing the dependencies

# @title
import os
import glob
import numpy as np
import pandas as pd
import weka.core.jvm as jvm
from weka.core import converters
import matplotlib.pyplot as plt

# @title
data_dir = 'data'

# @title
#!rm -r weka
#!rm -r data

# @title
#jvm.stop()
jvm.start(packages=True)

DEBUG:weka.core.jvm:Adding bundled jars
DEBUG:weka.core.jvm:Classpath=['/usr/local/lib/python3.10/dist-
packages/javabridge/jars/rhino-1.7R4.jar',
'/usr/local/lib/python3.10/dist-packages/javabridge/jars/runnablequeue
.jar',
'/usr/local/lib/python3.10/dist-packages/javabridge/jars/cpython.jar',
'/usr/local/lib/python3.10/dist-packages/weka/lib/mtj.jar',
'/usr/local/lib/python3.10/dist-packages/weka/lib/core.jar',
'/usr/local/lib/python3.10/dist-packages/weka/lib/weka.jar',
'/usr/local/lib/python3.10/dist-packages/weka/lib/arpack_combined.jar'
, '/usr/local/lib/python3.10/dist-packages/weka/lib/python-weka-
wrapper.jar']
DEBUG:weka.core.jvm:MaxHeapSize=default
DEBUG:weka.core.jvm:Package support enabled

# @title
# Preparing Datasets
if not os.path.exists(data_dir):
    !mkdir $data_dir
    for file in ['airline.arff', 'breast-cancer.arff', 'contact-
lenses.arff', 'cpu.arff', 'cpu.with.vendor.arff', 'credit-g.arff',
'diabetes.arff', 'glass.arff', 'hypothyroid.arff', 'ionosphere.arff',
'iris.2D.arff', 'iris.arff', 'labor.arff', 'segment-challenge.arff',
'segment-test.arff', 'soybean.arff', 'supermarket.arff',
'unbalanced.arff', 'vote.arff', 'weather.nominal.arff',
'weather.numeric.arff',]:
        url =
'https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/' + file
        !wget -P $data_dir $url
    loader =
converters.Loader(classname="weka.core.converters.ArffLoader")
    saver =
converters.Saver(classname="weka.core.converters.CSVSaver")
    for file in glob.glob(os.path.join(data_dir, '*.arff')):
        dataset = loader.load_file(file)
        filename, file_extension = os.path.splitext(file)
        saver.save_file(dataset, filename + '.csv')
    !wget -P $data_dir https://raw.githubusercontent.com/Rytuo/ITMO-
CT/master/Others/AdvancedML/data/OpenML/data/1438.arff
    !rm -r weka
```

```
--2023-09-29 14:14:08--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/airline.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2357 (2.3K) [text/plain]
Saving to: ‘data/airline.arff’

airline.arff        100%[===================>]   2.30K  --.-KB/s    in
0s

2023-09-29 14:14:10 (977 MB/s) - ‘data/airline.arff’ saved [2357/2357]

--2023-09-29 14:14:10--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/breast-cancer.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 29418 (29K) [text/plain]
Saving to: ‘data/breast-cancer.arff’

breast-cancer.arff  100%[===================>]  28.73K   180KB/s    in
0.2s

2023-09-29 14:14:11 (180 KB/s) - ‘data/breast-cancer.arff’ saved
[29418/29418]

--2023-09-29 14:14:11--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/contact-lenses.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2890 (2.8K) [text/plain]
Saving to: ‘data/contact-lenses.arff’
```

```
contact-lenses.arff 100%[===================>]   2.82K  --.-KB/s    in
0s

2023-09-29 14:14:11 (1.38 GB/s) - ‘data/contact-lenses.arff’ saved
[2890/2890]

--2023-09-29 14:14:12--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/cpu.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 5561 (5.4K) [text/plain]
Saving to: ‘data/cpu.arff’

cpu.arff            100%[===================>]   5.43K  --.-KB/s    in
0s

2023-09-29 14:14:12 (14.9 MB/s) - ‘data/cpu.arff’ saved [5561/5561]

--2023-09-29 14:14:12--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/cpu.with.vendor.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 6960 (6.8K) [text/plain]
Saving to: ‘data/cpu.with.vendor.arff’

cpu.with.vendor.arf 100%[===================>]   6.80K  --.-KB/s    in
0s

2023-09-29 14:14:13 (97.1 MB/s) - ‘data/cpu.with.vendor.arff’ saved
[6960/6960]

--2023-09-29 14:14:13--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/credit-g.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 162270 (158K) [text/plain]
Saving to: ‘data/credit-g.arff’
```

```
credit-g.arff       100%[===================>] 158.47K   329KB/s    in
0.5s

2023-09-29 14:14:14 (329 KB/s) - ‘data/credit-g.arff’ saved
[162270/162270]

--2023-09-29 14:14:14--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/diabetes.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 37443 (37K) [text/plain]
Saving to: ‘data/diabetes.arff’

diabetes.arff       100%[===================>]  36.57K   230KB/s    in
0.2s

2023-09-29 14:14:15 (230 KB/s) - ‘data/diabetes.arff’ saved
[37443/37443]

--2023-09-29 14:14:15--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/glass.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 17850 (17K) [text/plain]
Saving to: ‘data/glass.arff’

glass.arff          100%[===================>]  17.43K   110KB/s    in
0.2s

2023-09-29 14:14:16 (110 KB/s) - ‘data/glass.arff’ saved [17850/17850]

--2023-09-29 14:14:16--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/hypothyroid.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 310897 (304K) [text/plain]
Saving to: ‘data/hypothyroid.arff’
```

```
hypothyroid.arff    100%[===================>] 303.61K   477KB/s    in
0.6s

2023-09-29 14:14:17 (477 KB/s) - ‘data/hypothyroid.arff’ saved
[310897/310897]

--2023-09-29 14:14:18--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/ionosphere.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 80487 (79K) [text/plain]
Saving to: ‘data/ionosphere.arff’

ionosphere.arff     100%[===================>]  78.60K   165KB/s    in
0.5s

2023-09-29 14:14:19 (165 KB/s) - ‘data/ionosphere.arff’ saved
[80487/80487]

--2023-09-29 14:14:19--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/iris.2D.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3492 (3.4K) [text/plain]
Saving to: ‘data/iris.2D.arff’

iris.2D.arff        100%[===================>]   3.41K  --.-KB/s    in
0s

2023-09-29 14:14:19 (63.5 MB/s) - ‘data/iris.2D.arff’ saved
[3492/3492]

--2023-09-29 14:14:20--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/iris.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 7486 (7.3K) [text/plain]
```

```
Saving to: ‘data/iris.arff’

iris.arff           100%[===================>]   7.31K  --.-KB/s    in
0s

2023-09-29 14:14:20 (68.0 MB/s) - ‘data/iris.arff’ saved [7486/7486]

--2023-09-29 14:14:20--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/labor.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 8279 (8.1K) [text/plain]
Saving to: ‘data/labor.arff’

labor.arff          100%[===================>]   8.08K  --.-KB/s    in
0s

2023-09-29 14:14:21 (99.4 MB/s) - ‘data/labor.arff’ saved [8279/8279]

--2023-09-29 14:14:21--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/segment-challenge.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 200410 (196K) [text/plain]
Saving to: ‘data/segment-challenge.arff’

segment-challenge.a 100%[===================>] 195.71K   308KB/s    in
0.6s

2023-09-29 14:14:23 (308 KB/s) - ‘data/segment-challenge.arff’ saved
[200410/200410]

--2023-09-29 14:14:23--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/segment-test.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 109984 (107K) [text/plain]
Saving to: ‘data/segment-test.arff’
```

```
segment-test.arff   100%[===================>] 107.41K   225KB/s    in
0.5s

2023-09-29 14:14:24 (225 KB/s) - ‘data/segment-test.arff’ saved
[109984/109984]

--2023-09-29 14:14:24--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/soybean.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 202935 (198K) [text/plain]
Saving to: ‘data/soybean.arff’

soybean.arff        100%[===================>] 198.18K   311KB/s    in
0.6s

2023-09-29 14:14:25 (311 KB/s) - ‘data/soybean.arff’ saved
[202935/202935]

--2023-09-29 14:14:25--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/supermarket.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2025871 (1.9M) [text/plain]
Saving to: ‘data/supermarket.arff’

supermarket.arff    100%[===================>]   1.93M  1.73MB/s    in
1.1s

2023-09-29 14:14:27 (1.73 MB/s) - ‘data/supermarket.arff’ saved
[2025871/2025871]

--2023-09-29 14:14:27--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/unbalanced.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 186360 (182K) [text/plain]
```

```
Saving to: ‘data/unbalanced.arff’

unbalanced.arff     100%[===================>] 181.99K   286KB/s    in
0.6s

2023-09-29 14:14:29 (286 KB/s) - ‘data/unbalanced.arff’ saved
[186360/186360]

--2023-09-29 14:14:29--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/vote.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 40261 (39K) [text/plain]
Saving to: ‘data/vote.arff’

vote.arff           100%[===================>]  39.32K   247KB/s    in
0.2s

2023-09-29 14:14:30 (247 KB/s) - ‘data/vote.arff’ saved [40261/40261]

--2023-09-29 14:14:30--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/weather.nominal.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 587 [text/plain]
Saving to: ‘data/weather.nominal.arff’

weather.nominal.arf 100%[===================>]     587  --.-KB/s    in
0s

2023-09-29 14:14:30 (180 MB/s) - ‘data/weather.nominal.arff’ saved
[587/587]

--2023-09-29 14:14:30--
https://git.cms.waikato.ac.nz/weka/weka/-/raw/main/trunk/wekadocs/
data/weather.numeric.arff
Resolving git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)...
130.217.218.43
Connecting to git.cms.waikato.ac.nz (git.cms.waikato.ac.nz)|
130.217.218.43|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 495 [text/plain]
```

```
Saving to: ‘data/weather.numeric.arff’

weather.numeric.arf 100%[===================>]     495  --.-KB/s    in 
0s

2023-09-29 14:14:31 (156 MB/s) - ‘data/weather.numeric.arff’ saved 
[495/495]

--2023-09-29 14:14:32--  https://raw.githubusercontent.com/Rytuo/ITMO-
CT/master/Others/AdvancedML/data/OpenML/data/1438.arff
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 
185.199.109.133, 185.199.111.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|
185.199.109.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 25381 (25K) [text/plain]
Saving to: ‘data/1438.arff’

1438.arff           100%[===================>]  24.79K  --.-KB/s    in 
0.008s

2023-09-29 14:14:33 (3.00 MB/s) - ‘data/1438.arff’ saved [25381/25381]

rm: cannot remove 'weka': No such file or directory
```

Reading data

```
# Dataset names
# "airline", "breast-cancer", "contact-lenses", "cpu", 
"cpu.with.vendor", "credit-g", "diabetes", "glass", "hypothyroid", 
"ionosphere", "iris.2D", "iris", "labor", "segment-challenge", 
"segment-test", "soybean", "supermarket", "unbalanced", "vote", 
"weather.nominal", "weather.numeric"

# Read CSV dataset
pd.read_csv("data/weather.numeric.csv")

     outlook  temperature  humidity  windy play
0      sunny           85        85  False   no
1      sunny           80        90   True   no
2   overcast           83        86  False  yes
3      rainy           70        96  False  yes
4      rainy           68        80  False  yes
5      rainy           65        70   True   no
6   overcast           64        65   True  yes
7      sunny           72        95  False   no
8      sunny           69        70  False  yes
9      rainy           75        80  False  yes
10     sunny           75        70   True  yes
11  overcast           72        90   True  yes
```

```
12  overcast          81        75  False  yes
13     rainy          71        91   True   no
```

```
# Read Arff dataset
loader.load_file("data/weather.numeric.arff")
```

```
@relation weather

@attribute outlook {sunny,overcast,rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {TRUE,FALSE}
@attribute play {yes,no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Task 1.2.1 Read *iris* data.

Data Selection & Manipulation

- Pandas. 10 Minutes Guide
  - Basic data structures in pandas
  - Object creation
  - Viewing data
  - Selection
  - ~Missing data~
  - Operations
  - ~Merge~
  - Grouping
  - ~Reshaping~
  - ~Time series~
  - ~Categoricals~
  - ~Plotting~
  - ~Importing and exporting data~

    - ~Gotchas~
- Weka
    - Dataset Subsets

Task 1.2.2 Apply the given tutorials on weather dataframe.

Plotting

- Pandas Plotting
    - Basic plotting: plot
    - ~Other plots~
    - ~Plotting with missing data~
    - ~Plotting tools~
    - ~Plot formatting~
    - ~Plotting directly with Matplotlib~
    - ~Plotting backends~

Task 1.2.3 Apply the given tutorials on weather dataframe

Functional Programming (Don't Loop!)

A clean code principle in data manipulation is to never loop or iteratire on rows/columns. Always use mapping methods to transform the data.

```
# Create a DataFrame
df = pd.DataFrame({
    "name": ["Alice", "Bob", "Carol"],
    "age": [1, 5, 10]
  })

df

# Column-wise

# Define a mapping function
def square_age(age):
  return age * age

# Apply the mapping function to a dataframe's column
df["age"].apply(square_age)

0      1
1     25
2    100
Name: age, dtype: int64
```

```
# Lambda function
df['age'].apply(lambda x: x * x)

0      1
1     25
2    100
Name: age, dtype: int64

# Row-wise
def add_one(row):
  row["age"] = row["age"]+1
  return row

df.apply(add_one, axis=1)

    name  age
0  Alice    3
1    Bob    7
2  Carol   12

# Element-wise
df.applymap(lambda x: type(x))

            name            age
0  <class 'str'>  <class 'int'>
1  <class 'str'>  <class 'int'>
2  <class 'str'>  <class 'int'>
```

Task 1.2.4 Apply the given tutorial on weather dataset.

Task 1.2.5 On the age dataframe. Map names to the length of their strings. Then use `apply` function twice to compute the total sum of ages and names lengths.

Task 1.2.6 Solve *task 5* again but using `df.sum()` in place of `apply`.

1.3 Fielded Applications

Task 1.3.1 Select a favorite domain of your choice, like fashion or sports, and search for a use-case utilizing data mining in it.

Answer

1.4 The Data Mining Process

Task 1.4.1 Search for a tutorial for each step of the data-mining process.

Recall the cycle is: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment.

Answer

1.6 Generalization as Search

Task 1.6.1 Think of of model fitting as the problem of searching all possible curves separating the two features. Guess how can we ensure the search process always terminates. Guess some scenarios in which the search process yields unoptimal model.

Answer

1.7 Data mining and Ethics

Task 1.7.1 Consider a scenario in which learning from data concluded females are less likely to accommodate a loan, and as a result a girl's chances are less. Should we follow the data? Why?

Answer

Project. Phase 1

- Select some kaggle notebooks and a dataset.